Technical Progress Report 2/1/95 – 4/30/95
Construction of a Connectionist Network Supercomputer
University of California, Berkeley
ONR URI Grant No. N00014-92-J-1617

1 Abstract

This report presents a summary of the technical status for the period 2/1/95–4/30/95. In this period we passed a major milestone in the development of connectionist network computers—we completed the design and fabrication of the T0 vector microprocessor. We also continued to make significant progress in system and application software for these systems, in porting neural network algorithms, and in the application of analog auditory preprocessors to speech recognition.

2 Technical Status

2.1 Hardware Development

Vector Processors. This quarter we finished the layout and verification of the vector microprocessor, T0. Completion of this chip design represents a major milestone towards the development of a connectionist network supercomputer. T0 is the computational core of our future connectionist network computers. The internal details of the T0 chip have been reported elsewhere.

The chip was fabricated on the Hewlett Packard 0.8μm process (CMOS26B), managed through MOSIS. We used a conservative set of geometric design rules, resulting in an effective process resolution of 1.0μm. Three fabricated wafers were delivered several weeks ago and we subsequently tested them at Digital Testing Services in Santa Clara. The initially test results are very encouraging. The limitations of wafer probing preclude high-speed test, however, these tests do give us a good idea of functional correctness. In this case, wafer probe test covers more than 80% of the active circuitry of the design. Exactly 40% of the fabricated die passed the wafer probe tests. This percentage is very good for a die of the size of T0. A more usual number would be in the range of 15%–30%. We attribute this good yield to our careful design style and our use of conservative geometric design rules.

Wafer probe testing and subsequent initial circuit board testing exposed a problem in the design. The chip uses an unusually high amount of quiescent power. We are currently searching for the source of the problem. Besides running hotter than expected, this problem will not limit the usefulness of the chip in systems, and will be corrected on later versions.

Vector Processing Systems. The first use of the T0 chip in a system will be as part of an add-on processor board for a SUN workstation. In previous progress reports we reported that we have designed a custom SBus board, called SPERT, to hold a T0 chip. This circuit board design is unique in that the T0 chip is wire-bonded directly to the board, without the usual chip package. As outlined in previous reports, there are several advantages to this approach. However, because the process is somewhat unusual, we have had to work closely with local fabricators as they refined the steps necessary to meet our specifications. We have received shipment on prototype quantities of printed circuit boards complete with mounted T0 chips. We are currently evaluating these boards.

The SPERT boards will be tested using a test fixture we developed specifically for the purpose. We have drawn on our success using elastomeric connector technology from our experimental high-speed signally chips. The test fixture allows us to perform full-speed chip and board tests prior to populating the board with memory and other support chips.

2.2 Software and Applications

Speech application. This quarter we focused attention on detailed implementation of back-propagation training of multi-layer perceptrons (MLPs) in the SPERT software/hardware environment. Although the SPERT board is not yet ready for use, we have a complete software environment, including a T0 chip simulator, running on workstations.

The initial results are encouraging but leave room for improvement. The routines will be optimized once the SPERT hardware comes online. Below are performance numbers for training speed given in millions of connection updates per second (MCUPS), at 33MHz clock rate (the low-end projected clock speed for the SPERT board). Note that these timings were generated using just a bare net, with no database functionality included. Real training runs will be slightly slower initially due to database overhead, although there are some simple optimizations left that may compensate, particularly in the case of small nets.

Net size (input x hidden x output)	Speed (MCUPS)
64 x 64 x 32	16.7
153 x 192 x 2	39.3
153 x 192 x 56	38.9
153 x 200 x 56	38.0
162 x 1000 x 61	50.7
512 x 512 x 61	63.9
512 x 512 x 512	54.7

For comparison, these numbers are faster than a four board RAP for large nets, and much faster than any sized RAP system for small nets. Our fastest workstation for online training (SGI Indigo) gets $\leq$2 MCUPs on small nets, depending on net size.

High-level software. The major advance of this period was the release of Sather 1.0.5. This is the first general release to incorporate the parallel Sather constructs and represents a major step forward. All future releases will also incorporate pSather. There has also been good progress on the connectionist simulator, ICSIM. Ben Gomes, the principal designer of ICSIM, successfully passed his qualifying exam and is proceeding with implementation.

2.3 Analog VLSI pre-processors.

As described in the previous progress report, we recently constructed a three-chip sound pre-processing system, using three copies of an enhanced version of our 128-channel spectral-shape auditory pre-processor. A major focus of the past three months has been using this new system as a pre-processor in speech recognition systems.

Last year, we performed similar experiments for several months, using a single-chip system. The parameters of the chip were tuned to produce a periodicity-based spectral representation. Using a 200-speaker, isolated-word, telephone quality database, with 13 words (1-9, "oh", "zero", "yes", "no"), we evaluated the performance of this single-chip system, in conjunction with a traditional Hidden-Markov-Model based speech recognition system (the HTK package by Entropic Software).

This evaluation resulted in a final error rate for the task of about 13%—encouraging for an initial effort, but weak in comparison with the 2% error rates achieved by neural-network-based recognition systems, used in conjunction with standard software-based front ends.

We are currently pursuing two interrelated avenues to improve these performance figures:

1. Our new 3-chip systems lets us compute several different representations of the sound simultaneously. By combining different representations, each of which is specialized for different features of speech sounds, we expect to improve performance over our earlier single-chip experiments. In addition to using the raw representations from the chip, we also intend to explore secondary representations of sounds that are computed from the chip representations: such representations code parameters such as pitch and FM/AM modulations.

2. Our new experiments will use a neural-network-based Hidden-Markov-Model speech recognition system. We expect the the discriminant training of the back-propagation algorithm to extract more information from our auditory representations than the non-discriminant learning methods of the HTK speech recognizer.

Current work over the past three month has involved building software support for doing speech recognition experiments with multiple representations, using neural-network tools. Early experiments, using two representations (a spectral-shape representation and a temporal onset representation) in conjunction with a 200-hidden-unit neural-network, showed promising results (error rates in the 6–7% range, for the task described in the second

paragraph). We are currently doing expanded experiments with the goal of attaining error performance competitive with software-based front-ends, and will report complete results in the next progress report.

3 Recent Publications

Lazzaro, J. and Wawrzynek, J., "A Multi-Sender Asynchronous Extension to the AER protocol," 16th Conference on Advanced Research in VLSI, IEEE Computer Society Press, 1995.

Lazzaro, J. and Wawrzynek, J., Poster presentation at "Machines that Learn" workshop, Snowbird, Utah, April 4-7. "Efficient Systems for Auditory Scene Analysis".

Philippsen, M. and Heinz, E. A., "Automatic Synchronization Elimination in Synchronous FORALLs," Frontiers '95:The Fifth Symposium on the Frontiers of Massively Parallel Computation, February 6–9, Mc Lean, VA, pp 350–357, 1995.